Proceedings of HLT-NAACL 2003
Short Papers , pp. 13-15
Edmonton, May-June 2003

Story Link Detection and New Event Detection are Asymmetric

Francine Chen
PARC
3333 Coyote Hill Rd
Palo Alto, CA 94304
fchen@parc.com

Ayman Farahat
PARC
3333 Coyote Hill Rd
Palo Alto, CA 94304
farahat@parc.com

Thorsten Brants
PARC
3333 Coyote Hill Rd
Palo Alto, CA 94304
thorsten@brants.net

Abstract

Story link detection has been regarded as a core technology for other Topic Detection and Tracking tasks such as new event detection. In this paper we analyze story link detection and new event detection in a retrieval framework and examine the effect of a number of techniques, including part of speech tagging, new similarity measures, and an expanded stop list, on the performance of the two detection tasks. We present experimental results that show that the utility of the techniques on the two tasks differs, as is consistent with our analysis.

1 Introduction

Topic Detection and Tracking (TDT) research is sponsored by the DARPA TIDES program. The research has five tasks related to organizing streams of data such as newswire and broadcast news (Wayne, 2000). A link detection (LNK) system detects whether two stories are "linked", or discuss the same event. A story about a plane crash and another story about the funeral of the crash victims are considered to be linked. In contrast, a story about hurricane Andrew and a story about hurricane Agnes are not linked because they are two different events. A new event detection (NED) system detects when a story discusses a previously unseen event. Link detection is considered to be a core technology for new event detection and the other tasks.

Several groups are performing research on the TDT tasks of link detection and new event detection (e.g., (Carbonell et al., 2001) (Allan et al., 2000)). In this paper, we compare the link detection and new event detection tasks in an information retrieval framework, examining the criteria for improving a NED system based on a LNK system, and give specific directions for improving each system separately. We also investigate the utility of a number of techniques for improving the systems.

2 Common Processing and Models

The Link Detection and New Event Detection systems that we developed for TDT2002 share many processing steps in common. This includes preprocessing to tokenize the data, recognize abbreviations, normalize abbreviations, remove stop-words, replace spelled-out numbers by digits, add part-of-speech tags, replace the tokens by their stems, and then generating term-frequency vectors. Document frequency counts are incrementally updated as new sources of stories are presented to the system. Additionally, separate source-specific counts are used, so that, for example, the term frequencies for the New York Times are computed separately from stories from CNN. The source-specific, incremental, document frequency counts are used to compute a TF-IDF term vector for each story. Stories are compared using either the cosine distance $sim(d_1, d_2) = \frac{\sum_t f(t,d_1) \cdot f(t,d_2)}{\sqrt{\sum_t f(t,d_1)^2 \cdot \sum_t f(t,d_2)^2}}$ or Hellinger distance $sim(d_1, d_2) = \sum_t \sqrt{\frac{f(d_1,t)}{\sum_t f(d_1,t)} \cdot \frac{f(d_2,t)}{\sum_t f(d_2,t)}}$ for terms t in documents d_1 and d_2. To help compensate for stylistic differences between various sources, e.g., news paper vs. broadcast news, translation errors, and automatic speech recognition errors (Allan et al., 1999), we subtract the average observed similarity values, in similar spirit to the use of thresholds conditioned on the sources (Carbonell et al., 2001)

3 New Event Detection

In order to decide whether a new document d describes a new event, it is compared to all previous documents and the document d^* with highest similarity is identified. If the score $score(d) = 1 - sim(d, d^*)$ exceeds a thresh-

old θ_s, then there is no sufficiently similar previous document, and d is classified as a new event.

4 Link Detection

In order to decide whether a pair of stories d_1 and d_2 are linked, we compute the similarity between the two documents using the cosine and Hellinger metrics. The similarity metrics are combined using a support vector machine and the margin is used as a confidence measure that is thresholded.

5 Evaluation Metric

TDT system evaluation is based on the number of false alarms and misses produced by a system. In link detection, the system should detect linked story pairs; in new event detection, the system should detect new stories. A *detection cost*

$$C_{Det} = C_{miss} \cdot P_{miss} \cdot P_{tar} + C_{FA} \cdot P_{FA} \cdot P_{nontar}. \quad (1)$$

is computed where the costs C_{miss} and C_{FA} are set to 1 and 0.1, respectively. P_{miss} and P_{FA} are the computed miss and false alarm probabilities. P_{tar} and P_{nontar} are the a priori target and non-target probabilities, set to 0.02 and 0.98, respectively. The detection cost is normalized by dividing by $\min(C_{miss} \cdot P_{tar}, C_{FA} \cdot P_{nontar})$ so that a perfect system scores 0, and a random baseline scores 1. Equal weight is given to each topic by accumulating error probabilities separately for each topic and then averaged. The *minimum detection cost* is the decision cost when the decision threshold is set to the optimal confidence score.

6 Differences between LNK and NED

The conditions for false alarms and misses are reversed for the LNK and NED tasks. In the LNK task, incorrectly flagging two stories as being on the same event is considered a false alarm. In contrast, in the NED task, incorrectly flagging two stories as being on the same event will cause a true first story to be missed. Conversely, incorrectly labeling two stories that are on the same event as not linked is a miss, but for the NED task, incorrectly labeling two stories on the same event as not linked may result in a false alarm.

In this section, we analyze the utility of a number of techniques for the LNK and NED tasks in an information retrieval framework. The detection cost in Eqn. 1 assigns a higher cost to false alarms since $C_{miss} \cdot P_{tar} = 0.02$ and $C_{FA} \cdot P_{nontar} = 0.098$. A LNK system should minimize false alarms by identifying only linked stories, which results in high precision for LNK. In contrast, a NED system will minimize false alarms by identifying all stories that are linked, which translates to high recall for LNK. Based on this observation, we investigated a number of precision and recall enhancing techniques for the

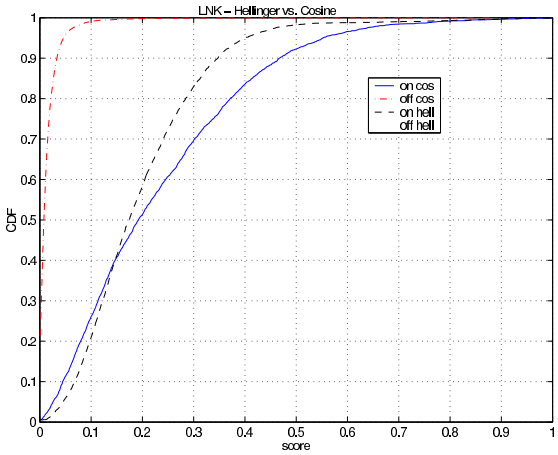

Figure 1: CDF for cosine and Hellinger similarity on the LNK task for on-topic and off-topic pairs.

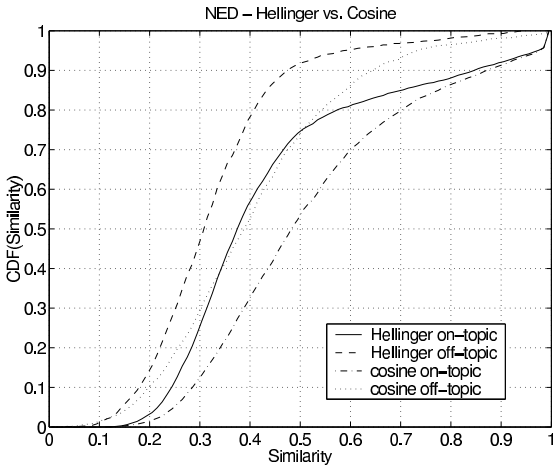

Figure 2: CDF for cosine and Hellinger similarity on the NED task for on-topic and off-topic pairs.

LNK and NED systems, namely, part-of-speech tagging, an expanded stoplist, and normalizing abbreviations and transforming spelled out numbers into numbers. We also investigated the use of different similarity measures.

6.1 Similarity Measures

The systems developed for TDT primarily use cosine similarity as the similarity measure. In work on text segmentation (Brants et al., 2002), better performance was observed with the Hellinger measure. Table 1 shows that for LNK, the system based on cosine similarity performed better; in contrast, for NED, the system based on Hellinger similarity performed better.

The LNK task requires high precision, which corresponds to a large separation between the on-topic and off-topic distributions, as shown for the cosine metric in Figure 1. The NED task requires high recall (low CDF

Table 1: Effect of different similarity measures on topic-weighted minimum normalized detection costs on the TDT 2002 dry run data.

System	Cosine	Hellinger	Change(%)
LNK	0.3180	0.3777	-0.0597(-18.8)
NED	0.7059	0.5873	+0.1186(+16.3)

Table 2: Effect of using part-of-speech on minimum normalized detection costs on the TDT 2002 dry run data.

System	− PoS	+ PoS	Change (%)
LNK	0.3180	0.3334	-0.0154 (−4.8%)
NED	0.6403	0.5873	+0.0530 (+8.3%)

values for on-topic). Figure 2, which is based on pairs that contain the current story and its most similar story in the story history, shows a greater separation in this region with the Hellinger metric. For example, at 10% recall, the Hellinger metric has 71% false alarm rate as compared to 75% for the cosine metric.

6.2 Part-of-Speech (PoS) Tagging

To reduce confusion among some word senses, we tagged the terms as one of five categories: adjective, noun, proper nouns, verb, or other, and then combined the stem and part-of-speech to create a "tagged term". For example, 'N_train' represents the term 'train' when used as a noun. The LNK and NED systems were tested using the tagged terms. Table 2 shows the opposite effect PoS tagging has on LNK and NED.

6.3 Stop Words

The broadcast news documents in the TDT collection have been transcribed using Automatic Speech Recognition (ASR). There are systematic differences between ASR and manually transcribed text. For example "30" will be spelled out as "thirty" and 'CNN" is represented as three separate tokens "C", "N", and "N". To handle these differences, an "ASR stoplist" was created by identifying terms with statistically different distributions in a parallel corpus of manually and automatically transcribed documents, the TDT2 corpus. Table 3 shows that use of an ASR stoplist on the topic-weighted minimum detection costs improves results for LNK but not for NED.

We also performed "enhanced preprocessing" to normalize abbreviations and transform spelled-out numbers into numerals, which improves both precision and recall. Table 3 shows that enhanced preprocessing exhibits worse performance than the ASR stoplist for Link Detection, but yields best results for New Event Detection.

Table 3: Effect of using an "ASR stoplist" and "enhanced preprocessing" for handling ASR differences on the TDT 2001 evaluation data.

ASRstop	No	Yes	No
Preproc	Std	Std	Enh
LNK	0.312	0.299 (+4.4%)	0.301 (+3.3%)
NED	0.606	0.641 (-5.5%)	0.587 (+3.1%)

7 Summary and Conclusions

We have presented a comparison of story link detection and new event detection in a retrieval framework, showing that the two tasks are asymmetric in the optimization of precision and recall. We performed experiments comparing the effect of several techniques on the performance of LNK and NED systems. Although many of the processing techniques used by our systems are the same, the results of our experiments indicate that some techniques affect the performance of LNK and NED systems differently. These differences may be due in part to the asymmetry in the tasks and the corresponding differences in whether improving precision or recall for the link task is more important.

8 Acknowledgments

We thank James Allan of UMass for suggesting that precision and recall may partially explain the asymmetry of LNK and NED.

References

James Allan, Hubert Jin, Martin Rajman, Charles Wayne, Dan Gildea, Victor Lavrenko, Rose Hoberman, and David Caputo. 1999. Topic-based novelty detection. Summer workshop final report, Center for Language and Speech Processing, Johns Hopkins University.

James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in TDT is hard. In *CIKM*, pages 374–381.

Thorsten Brants, Francine Chen, and Ioannis Tsochantaridis. 2002. Topic-based document segmentation with probabilistic latent semantic analysis. In *CIKM*, pages 211–218, McLean, VA.

Jaime Carbonell, Yiming Yang, Ralf Brown, Chun Jin, and Jian Zhang. 2001. Cmu tdt report. Slides at the TDT-2001 meeting, CMU.

Charles Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *LREC*, pages 1487–1494, Athens, Greece.